ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

AIR POLLUTION MODELING WITH MACHINE LEARNING TECHNIQUES

Pankaj Walde[1], Dr. Rajni Kant[2] and Shailendra Bommanwar[3]

[1]Research Scholar, Department of Mining Engineering, BIT, Ballarpur (MS), India

[2]Principal, Ballarpur Institute of Technology, Ballarpur, Distt. - Chandrapur (MS) -442901.

[3]Assistant Professor, Department of Mining Engineering, Ballarpur Institute of Technology, Ballarpur, Dist. - Chandrapur (MS) -442901

ABSTRACT

In the present scenario, the rises in population leading to the rise in demand of thesociety daily with time. Hence the consequences of our action are leading to an increase in air pollution and a decrease in the air quality index which is a severe and critical situation as it has an impact on human beings, flora and fauna and historical buildings in and around mining areas. Environmental concern is also increasing with time due to the release of ozone, sulphur dioxide and many other harmful gases which lead to the depletion of the ozone layer and this led to melting of icecaps lead to a rise in water level in the sea and an increase in the temperature. The environmental impact of coal mining cannot be ignored but, to some extent, isunavoidable. Most major mining activities contribute directly or indirectly to air pollution (Kumar et al., 1994; CMRI, 1998). Sources of air pollution in the coal mining areas generally include drilling, blasting, overburden loading and unloading, coal loading and unloading, haul roads, transport roads, stockyards, exposed overburden dumps, coal handling plant, exposed pit faces and workshop (CMRI, 1998). The major air pollutants produced are suspended particulate matter and respirable particulate matter which is in contrast to vehicular emissions where lead and gaseous pollutants are a major concern. An air quality index (AQI) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. Public health risks increase as the AQI rises. The Air Quality Index is based on the measurement of particulate matter (PM2.5 and PM10), Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2) and Carbon Monoxide (CO) emissions.

Keywords: Air Pollution, PM 2.5, P.M 10, Machine Learning Techniques.

1. INTRODUCTION

Air pollution is the introduction of particulates, biological molecules, or other harmful materials into the Earth's atmosphere, causing disease, death to humans, damage to other living organisms such as food crops, or damage to the natural or man-made environment. The United States Environmental Protection Agency (US-EPA) defined air pollution as one or more chemicals or substances in high enough concentrations in the air to harm humans, other animals, vegetation, or materials.

The WHO estimates that more than two million people die each year from causes directly attributable to air pollution.

Air pollution is the most dangerous form of pollution. The major air pollutants are SOX, NOX, CO, Particulate matter etc. It results from gaseous emission from industry, thermal power stations, domestic combustion etc. Due to air pollution, the composition of air is changing all over the world. Most of the gases and air pollutants are produced by burning fuels. The burning of coal produces carbon dioxide, sulfur dioxide etc. which are responsible for acid rain. Chlorofluorocarbons are widely used as propellants and as refrigerants that cause ozone depletion. The Taj Mahal in Agra is affected by the fumes emitted by the Mathura refinery. Reports

estimate that the monument would get defaced within twenty years because of the harmful effluents of the mission from the refinery. Theemission of greenhouse gases has led to climatic changes. The increase in pollution has resulted in global warming. Global warming is an average increase in the Earth's temperature due to the greenhouse effect as a result of both natural and human activity. The term climate is often used interchangeably with the term global warming. The ice caps in the Polar Regions have begun to melt fast. This has resulted in the

rise of the water level of the seas and oceans. Grass sprouting in Antarctica and snowfall in the desert of the United Arab Emirates are all the warning signals of global warming. These are caused by the Greenhouse Effect.

Primary pollutants are usually produced from a process such as ash from a volcanic eruption. Other examples include carbon monoxide gas from motor vehicle exhaust or sulfur dioxide released from factories. Secondary pollutantsare not emitted directly. Rather they form in the air when primary pollutants react or intersect. Ground-level ozone is a prominent example of secondary pollutants.

And the same way air pollutant sources can be classified according to a type of source, number, spatial distribution & emission type & anthropogenic sources. Natural air pollutant sources include dust that originates from natural sources, methane, a by-product of food digestion by animals, smoke and carbon monoxide from volcanic activity and forest fires, and decay of radioactive gases within the Earth's crust. Anthropogenic (or manmade) sources include smoke including soot particles from stationary sources such as stacks or chimneys of power plants, manufacturing facilities, and incinerators, mobile sources such as automobiles, controlled burning practices used in agriculture and forestry sectors. Particle pollution (also called particulate matter or PM) is the term for a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope.

Fine particulate matter (PM2.5) consisting of particles with a diameter of 2.5 μm or smaller, is an important pollutant among the criteria pollutants. The U.S. EPA defines PM10 as particulate matter with a diameter of 10micrometres collected with 50% efficiency by a PM10 sampling collection device. However, for convenience in this study, the term PM10 will be used to include all particles having an aerodynamic diameter of less than or equal to 10micrometres. PM10 is regulated as a specific type of "pollutant" because this size range are considered reparable. In other words, particles less than approximately 10micrometres can penetrate the lower respiratory tract. The particle size range between 0.1 and 10micrometre is especially important in air pollution studies. A major fraction of the particulate matter generated in some industrial sources is in this size range. As with PM10, EPA defines PM2.5 as particulate matter with a diameter of 2.5micrometres collected with 50% efficiency by a PM2.5 sampling collection device. However, for convenience, the term PM2.5 will be used to include all particles having an aerodynamic diameter of less than or equal to 2.5micrometres. Particles less than approximately 2.5micrometres are regulated as PM2.5. Air emission testing and air pollution control methods for PM2.5 particles are different than those for coarse and super coarse particles. PM2.5 particles settle quite slowly in the atmosphere relative to coarse and super coarse particles.

2. AIM OF THE STUDY

The goal of this project to Air pollution modeling with machine learning to provide a means of calculating ambient ground-level concentrations of an emitted substance giveninformation about the emissions and the nature of the atmosphere.

3. OBJECTIVE OF THE STUDY

- Selection of an area for air pollution modeling.
- Collection of micro-meteorological data for the duration of sampling.
- Understanding the relation of meteorological data.
- Modeling of air pollution using the above data

4. LITERATURE REVIEW

4.1 Different types of Air Pollution Models:

Modeling of pollutant dispersion is completed using mathematical algorithms. There are several basic mathematical algorithms in use Box model Gaussia model Eulerian model Lagrangian model

4.1.1 Box model algorithm the box model is the simplest of the modeling algorithms. It assumes the airshed in the shape of a box. The box model isrepresented using following equation –

$$\frac{d(CV)}{dt} = Q * A + u * C_{in} * W * H - u * C * W * H$$

Where, Q = pollutant emission rate per unit area

C = homogeneous species concentration within the airshed V = volume described by box

Cin = species concentration entering airshed A = horizontal area of box

u = wind speed normal to the box H = mixing height

Although useful, this model has limitations. It assumes the pollutant is homogeneous across theairshed, and it is used to estimate average pollutantconcentrations over very large area.

4.1.2 Gaussian model algorithm The Gaussian models are the most common mathematical models used forair dispersion. They are based upon the assumptionthat the pollutant will disperse according to the normal statistical distribution.

Gaussian distribution equation is given by

$$(x, y, z) = \frac{Q}{2\pi u \sigma_y \sigma_z} + [\{\exp(-\frac{(z-\square)^2}{2\sigma_z^2}) + \exp(-\frac{(z+\square)^2}{2\sigma_z^2})\} \{\exp(-\frac{y^2}{2\sigma_z^2})\}]$$

Where,

C (x,y,z) = Pollutant concentration as a function of downwind position (x, y, z)

Q = mass emission rate u = wind speed

σy = standard deviation of pollutant concentration in y (horizontal) direction

σz = standard deviation of pollutant concentration in z (vertical) direction y = distance in horizontal direction z = distance in vertical direction H = effective stackheight The Gaussian distribution determines the size of the plume downwind fromthe source.

The Gaussian distribution determines the size of the plume downwind from the source. A schematic representation of the Gaussian Plume is shown in Figure 3.1. The plume size is dependent on the stability of the atmosphere and the dispersion ofthe plume in the horizontal and vertical directions. These horizontal and vertical dispersion coefficients (σy and σz respectively) are merely the standard deviation from normal on the Gaussian distribution curve in the y and z directions. These dispersion coefficients, σy and σz, are functions of wind speed, cloud cover, and surface heating by the sun. The Gaussian distribution requires that the material in the plume be maintained.

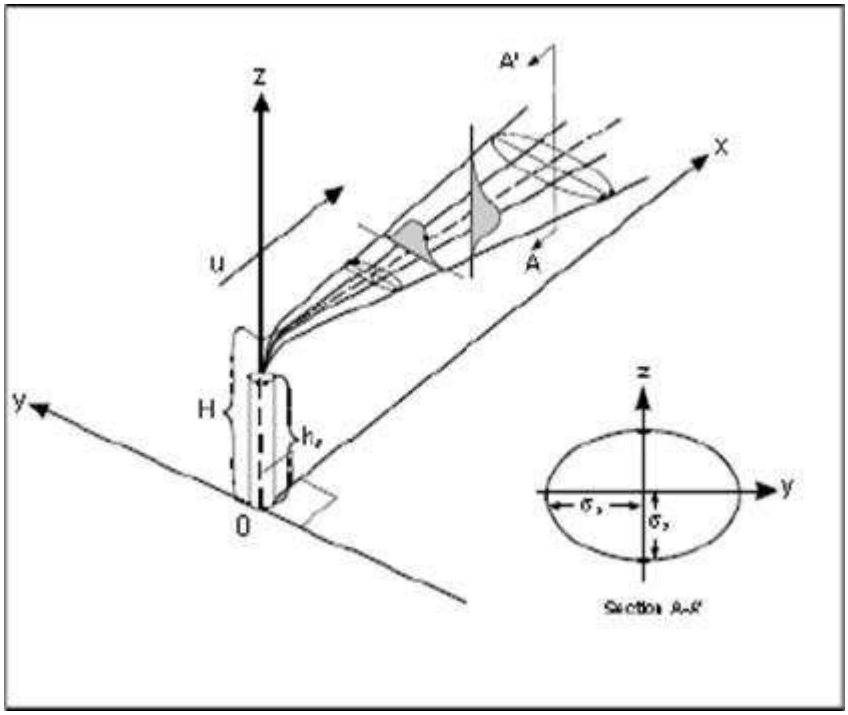


Fig.3.1 Schematic representation of Gaussian Plume

In order for a plume to be modeled using the Gaussian distribution the followingassumption must be made:

- The plume spread has a normal distribution
- The emission rate (Q) is constant and continuous
- Wind speed and direction is uniform
- Total reflection of the plume takes place at the surface
- The terrain is relatively flat, i.e., no crosswind barriers

4.1.3 Eulerian model algorithm Eulerian model solves a conservation of mass equation for a given pollutant. Equation follows the form:

$$\frac{\partial < c_i >}{\partial t} = -U * \Delta < c > -\Delta < c\, U' > + D\Delta^2 < > + < S >$$

$i \qquad i \qquad i \qquad i$

Where, U = Ū + U' U = wind field vector

U(x, y, z) Ū = average wind field vector U' = fluctuating wind field vector c = < c > + c' c

= pollutant

Concentration < c > = average pollutant concentrationc' = fluctuating pollutant concentration

D = molecular diffusivity

Si = source term

This equation can be difficult to solve because the advection term -Ū *Δ<ci>, is hyperbolic, the turbulent diffusion term is parabolic, and the source term is generallydefined by a set of differential equations. This type of equation can be computationally expensive to solve and requires some form of optimization in orderto reduce the solution time required.

4.1.4 Lagrangian model algorithm Lagrangian models predict pollutant dispersionbased on a shift in reference grid. This shifting reference grid is generally based on the prevailing wind direction, or vector, or the general direction of the dust plume movement. The Lagrangian model has the following form:

$$< c(r, t) > = \int_{-\infty}^{t} p(r, t \mid r', t')\, S(r', t')\, dr'\, dt'$$

Where, < c (r, t) > = average pollutant concentrationat location r at time t

S (r', t') = source emission term

p (r, t | r', t') = probability function that an air parcel is moving from location r' at timet' to location r

at time t

This mathematical model has limitations when its results are compared with actual measurements. This is due to the dynamic nature of the model. Measurements are generally made at stationary points, while the model predicts pollutant concentration based upon a moving reference grid.

4.1.5 Meteorological models: Meteorological models are developed for two purposes: To understand local, regional, or global meteorological phenomena; and To provide the meteorological input required by air pollution dispersion models. Numerical meteorological models can be divided into two groups:

1. Diagnostic models, i.e., models that interpolate and extrapolate availablemeteorological measurements and contain no time-tendency terms; and
2. Prognostic models, i.e., models with full time-dependent equations.

Machine learning is a major sub-field in computational intelligence (also called artificial intelligence). Its main objective is to use computational methods to extractinformation from data. Machine learning has a wide spectrum of applications including handwriting and speech recognition, robotics and computer games, natural language processing, brain-machine interface and so on. In the environmental sciences, machine learning methods have been heavily used in data processing, model emulation, weather and climate prediction, air quality forecasting, oceanographic and hydrological forecasting (Hsieh, 2009).

Linear Regression:
The term regression is used when you try to find the relationship between variables.Linear regression is perhaps one of the most well-known and well understood.

Linear regression models can then be further separated into simple and multiple linear regressions. Simple regression refers to a model which maps a linear relationship between a singular output andinput.

An actual linear relationship can be represented as such:

$y_i = \beta_0 + \beta_1 X_i + \varepsilon$

y i (y_i) represents the actual output given the input (X_i)

ε represents the random error term. The random error term represents the residualerror

between an estimated relationship and the actual one. It is subject to variability.

Multiple linear regression is really similar to simple linear regression. However, it maps the relationship between multiple types of inputs andan output.

An estimate of the relationship can be given as:

$\hat{y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta n X_n$

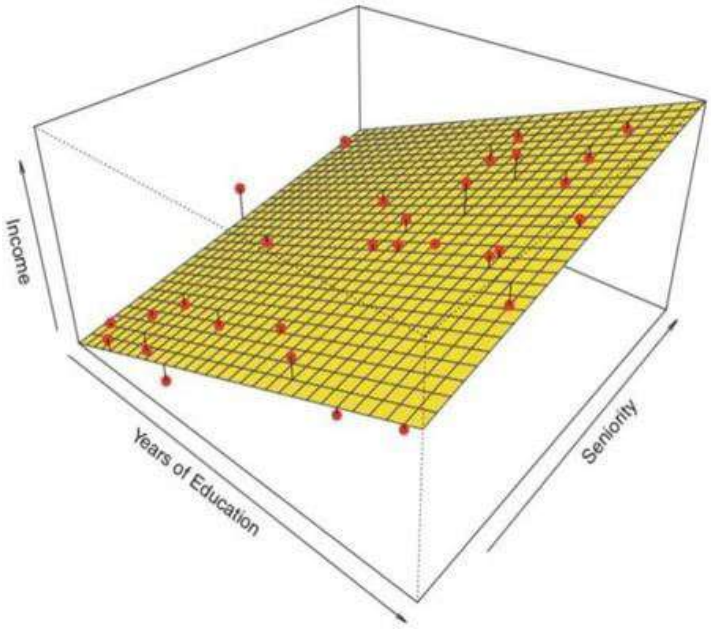
Where n represents subscript n

y hat sub i ($\hat{y}_i$) represents the estimated output given the input. (X_i)β_0 is a bias and β_1, … βn are the weights of the model.

X_{i1}, … X_in represents all of the individual input datapoints where X_{i1}represents a singulardatapoint of input type X_1.

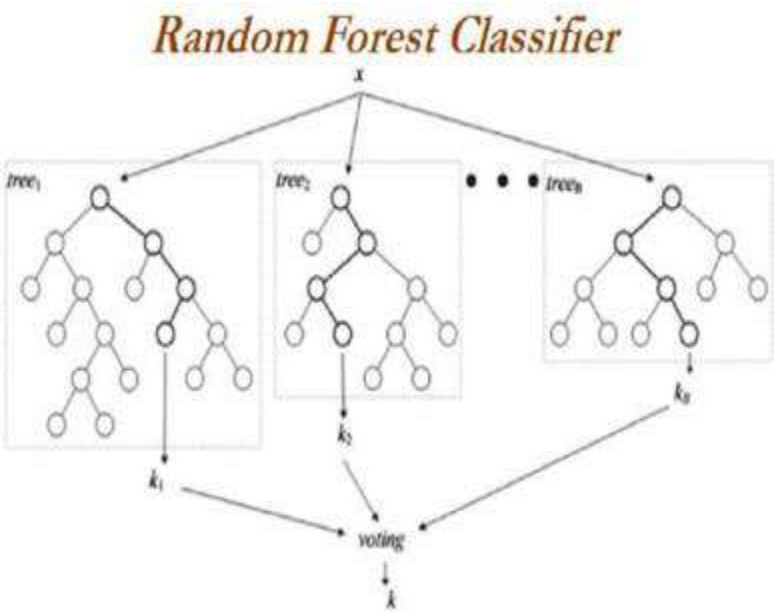
n is the number of input data types

A multiple linear regression equation can be graphed as an n-dimensional plane.



How Random Forest works

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at randomforest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



K-Nearest Neighbor:

K-nearest neighbor's algorithm (*k*-NN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951. It is used for classification and regression. In both cases, the input consists of the closest trainingexamples in data set. The output depends on whether k-NN is used for classificationor regression:

In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assignedto the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

Extreme Gradient Boosting or XGBoost Regression:

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. Regression predictive modeling problems involve predicting numerical value such as a dollar amount or a height. XGBoost can beused directly for regression predictive modeling.

Artificial Neural Network (ANN):

Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems vaguely inspired by the biological neural networks that constitute animal brains.

Neural Network

Neural network (NN) methods were originally developed from investigations into humanbrain function and they are adaptive systems that change as they learn (Hsieh and Tang, 1998). There are many types of NN models, the most common one is the multi-layer perceptron (MLP) NN model.

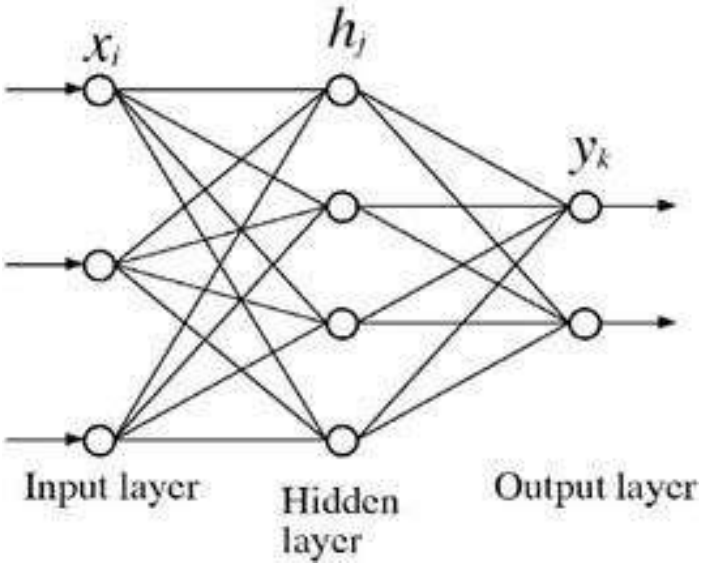


Fig: - NN model

The input variables xi are mapped to a layer of intermediate variables known as "hidden neuron" h_j by and then onto the output variables y_k by

$$h_j = f(\sum_i w_{ji} x_i + b_j),$$

$$y_k = g(\sum_j \beta_{kj} h_j + \beta_{k0})$$

Where f and g are "activation" functions in the hidden layer and the output layer, respectively. Normally can be the logistic sigmoidal or hyperbolic tangent function and g can be linear in NN models for regression. wji and βj are weight parameters and bj and βk0 are offset parameters. Their optimal values are learned by model training (Hsieh and Tang, 1998) where the mean squared error of the model output is minimized.

METHODOLOGY

All data loading, cleaning, processing, analyses, statistical modeling and visualization was carried out using python with many tools used from the packages like pandas, NumPy, seaborn and others. Temporal trends in the PM10 concentrations and corresponding meteorological and environmental predictor variables were estimated using the sklearn packages available and developed by goggle. ANN used the package which is built by Tensor Flow.

CASE STUDY: - Chandrapur (earlier known as *Chanda*) is a city and a municipal corporation in Chandrapur district, Maharashtra state, India. It is the district headquarters of Chandrapur district. Chandrapur is a fort city founded by Khandkya Ballal Sah, a Gond king of the 13th century. The city sits at the confluence of the Irai River and Zarpat River. The area around the city is rich in coal seams. Hence, Chandrapur is also known as the "black gold city". It is located in central India in the eastern part of Maharashtra state at 19.57°N latitude and 79.18°E longitude. The place is situatedat 189.90 meters above the mean sea level. The area of the city is about 70.02 km². The north-south length of the city is about 10.6 km, while the east-west length is about 7.6 km. The city slopes from the north to the south. Chandrapur lies at the confluence of the Irai and Zarpat rivers. The Irai river has a history of flooding. Flood marks are seen on the walls of the city. The Gaontideo Nala originates from the uplands of the Chandrapur Super Thermal Power Station. Chandrapur has a hot and dry climate. December is

the coldest month, with a minimum average temperature of 9 °C and a maximum average temperature of 23.2 °C. May is the hottest month with a mean maximum temperature of 43 °C and a mean minimum temperature of 28.2 °C.

As per provincial reports of Census of India, the population of Chandrapur city in 2011 was 3,20,379 of which males and female numbers were 1,64,085 and 1,56,294 respectively. Annual growth rate in population is 1.091% in 2011.

Chandrapur city	Total Population	Male	Female
Population	3,20,379	1,64,085	1,56,294

Demographic of Chandrapur city

The total number of registered motor vehicles till 2011 is 281764 as per RTPstatistic data. It is increasing at the rate of 10.7% annually. Data:

The data collected contain both particles and gaseous pollutants (PM2.5, PM10, NOx, SO2, O3 and CO). Other meteorological variables include wind speed, wind direction, rainfall, pressure, temperature andrelative humidity. The data is collected from the CPCB (www.cpcb.gov.in) and Meteoblue (www.meteoblue.com). There are two measuring stations in Chandrapur city. The data which is used here is from the CPCB station near the bus stand. The collected data is from July 2015 till March 2021. The data has 51,136 rows and 14 columns.

A range of meteorological, environmental and temporal variables was gathered to use as predictors for the PM10 measured in Chandrapur. Each of the meteorological and environmental variables is described in the below table.

Data Preparation:

The predictor variable selected for the ML modeling is PM10 and other dependent variables. The dependent variables are PM2.5, SO2, OZONE, RAIN, TEMPERATURE, PRESSURE, WIND SPEED AND WIND DIRECTION.

Predictor Variable	Variable Type	Source
PM2.5 NO2	Environmental	CPCB
SO2 CO	Environmental	CPCB
OZONE	Environmental	CPCB
Relative humidity	Environmental	CPCB
Temprature	Meteorological	CPCB
SpeedWind	Meteorological	CPCB
Direction Rain	Meteorological	CPCB

Measured at the same monitoring station in Chandrapur as the PM10 measurements. To provide a comparison of the daily and weekly cycles of PM10, air quality data from the Metoblue. Daily rainfalldata was fetched from the CPCB organization.

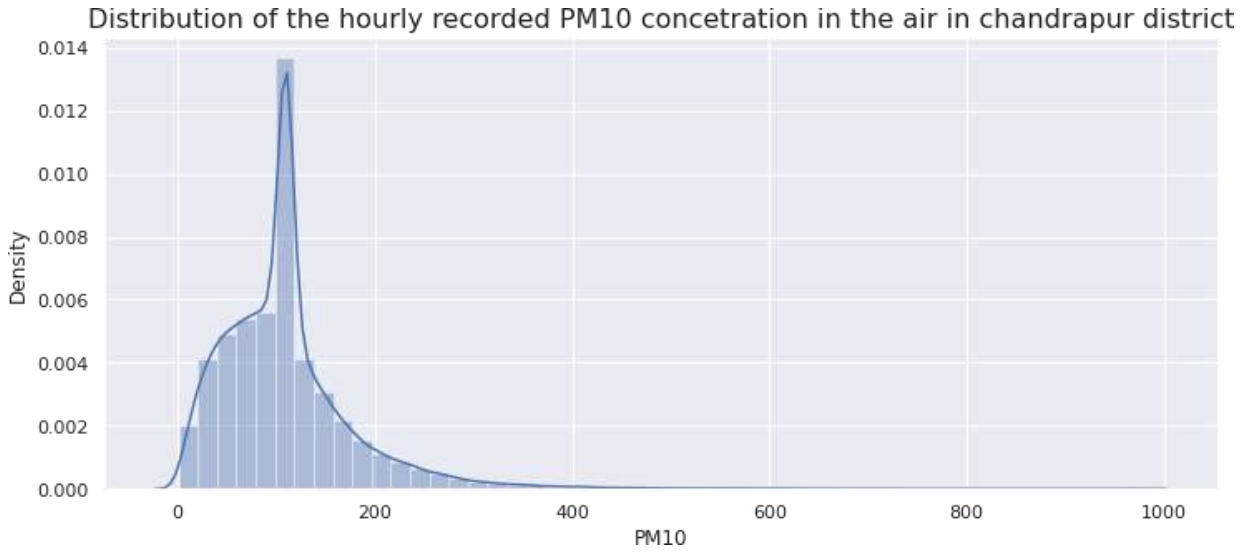
From Date	To Date 01-07-2015	PM2.5	PM10	NO2	SO2	CO	O3	RH	TEMP	WS	WD	RAIN	PRESSURE
	01-07-2015	22	46	2.34	3.44	0.66	5.49	69.3	23.8	2.2	239.83	1.6	1005.3
01-07-2015	01-07-2015	16	39	2.52	3.54	0.27	5.2	69.1	23.8	1.5	272.69	1.3	1004.9
01-07-2015	01-07-2015	18	44	2.57	3.43	0.31	5.37	70.9	23.7	1.2	281.26	0.5	1004.4
01-07-2015	01-07-2015	14	47	1.76	3.54	0.29	5.19	71.6	23.7	2	258.77	0	1003.8

01-07-2015	01-07-2015	15	51	1.9	3.61	0.23	4.8	72.1	23.7	1.9	173.73	0	1003.6
01-07-2015	01-07-2015	13	40	2.75	3.65	0.35	4.22	72.7	23.7	1.5	179.23	0	1003.8
01-07-2015	01-07-2015	24	56	3.12	3.72	0.75	4.73	72.1	23.6	1.5	254.61	0	1004.4
01-07-2015	01-07-2015	25	76	2.97	3.66	1.42	3.74	70.8	29.2	2	105.29	0	1005.2
01-07-2015	01-07-2015	41	109	2.51	5	1.37	0.68	69.7	39.4	1.9	161.79	0	1005.9
01-07-2015	01-07-2015	27	49	2.41	5.88	0.9	None	67.7	44	2	137.77	0	1006.2
01-07-2015	01-07-2015	27	49	2.28	2.88	0.54	4.72	64.4	37.9	1.7	185.84	0.1	1005.8
01-07-2015	01-07-2015	None	None	2.29	4.21	0.21	7.8	60.1	28.8	1.8	227.46	0.7	1005.3
01-07-2015	01-07-2015	28	51	1.75	5.09	0.5	5.72	60	26.9	2	213.52	0.7	1005.2
01-07-2015	01-07-2015	34	57	1.56	4.08	0.6	5.27	60.8	27.7	2.3	197.21	0.7	1005.1
01-07-2015	01-07-2015	26	41	1.1	3.93	0.62	5.77	60	28	2.7	190.63	0.9	1004.7
01-07-2015	01-07-2015	23	51	1.47	5.25	0.78	4.97	56.6	29.6	1.7	215.78	0.3	1004.4
01-07-2015	01-07-2015	31	66	2.33	4.16	0.42	6.91	57.4	27.8	3.8	290.97	0.3	1004.5
01-07-2015	01-07-2015	14	44	4.07	3.54	0.5	7.73	62	26.2	2.7	254.43	0	1004.9
01-07-2015	01-07-2015	11	44	2.27	3.45	0.44	5.28	60.2	22.9	2.3	270.66	0	1005.4
01-07-2015	01-07-2015	8	32	2.66	3.45	0.56	5.35	61.5	23.2	1	207.2	0	1006
01-07-2015	01-07-2015	13	50	2.61	3.41	0.7	5.05	63.7	23.4	1.5	125.22	0	1006.6
01-07-2015	01-07-2015	24	67	2.47	3.31	0.89	4.59	65.9	23.4	1.4	182.2	0	1006.6
01-07-2015	01-07-2015	18	57	1.49	3.32	0.43	5.05	66.7	23.4	1.4	194.14	0	1006.7
01-07-2015	02-07-2015	7	36	1.65	3.34	0.37	4.85	67.2	23.5	1.7	163.56	0	1006.5
02-07-2015	02-07-2015	12	45	2.31	3.3	0.25	4.99	66.9	23.6	1.7	179.45	0	1005.8

The figure will show the data after being processed with the above code

date	pm25	pm10	no2	so2	co	ozone	rh	temp	ws	wd	rain	pressure	year	month	day	hour	wdd	dewp
2015-07-01 00:00:00	22	46	2.34	3.44	0.66	5.49	69.3	23.79	2.18	239.83	1.6	1005.3	2015	7	1	0	WSW	17.65
2015-07-01 01:00:00	16	39	2.52	3.54	0.27	5.2	69.1	23.78	1.51	272.69	1.3	1004.9	2015	7	1	1	W	17.6
2015-07-01 02:00:00	18	44	2.57	3.43	0.31	5.37	70.85	23.68	1.18	281.26	0.5	1004.4	2015	7	1	2	WNW	17.85
2015-07-01 03:00:00	14	47	1.76	3.54	0.29	5.19	71.6	23.71	1.96	258.77	0	1003.8	2015	7	1	3	W	18.03
2015-07-01 04:00:00	15	51	1.9	3.61	0.23	4.8	72.13	23.66	1.86	173.73	0	1003.6	2015	7	1	4	S	18.09
2015-07-01 05:00:00	13	40	2.75	3.65	0.35	4.22	72.67	23.67	1.5	179.23	0	1003.8	2015	7	1	5	S	18.2
2015-07-01 06:00:00	24	56	3.12	3.72	0.75	4.73	72.12	23.64	1.51	254.61	0	1004.4	2015	7	1	6	WSW	18.06
2015-07-01 07:00:00	25	76	2.97	3.66	1.42	3.74	70.77	29.2	1.99	105.29	0	1005.2	2015	7	1	7	ESE	23.35
2015-07-01 08:00:00	41	109	2.51	5	1.37	0.68	69.74	39.44	1.9	161.79	0	1005.9	2015	7	1	8	SSE	33.39
2015-07-01 09:00:00	27	49	2.41	5.88	0.9	16.48	67.69	43.99	1.99	137.77	0	1006.2	2015	7	1	9	SE	37.53
2015-07-01 10:00:00	27	49	2.28	2.88	0.54	4.72	64.42	37.86	1.67	185.84	0.1	1005.8	2015	7	1	10	S	30.74
2015-07-01 11:00:00	985	985	2.29	4.21	0.21	7.8	60.08	28.83	1.83	227.46	0.7	1005.3	2015	7	1	11	SW	20.85
2015-07-01 12:00:00	28	51	1.75	5.09	0.5	5.72	60.02	26.9	2.02	213.52	0.7	1005.2	2015	7	1	12	SSW	18.9
2015-07-01 13:00:00	34	57	1.56	4.08	0.6	5.27	60.84	27.74	2.31	197.21	0.7	1005.1	2015	7	1	13	SSW	19.91
2015-07-01 14:00:00	26	41	1.1	3.93	0.62	5.77	59.95	27.97	2.7	190.63	0.9	1004.7	2015	7	1	14	S	19.96
2015-07-01 15:00:00	23	51	1.47	5.25	0.78	4.97	56.61	29.59	1.68	215.78	0.3	1004.4	2015	7	1	15	SW	20.91
2015-07-01 16:00:00	31	66	2.33	4.16	0.42	6.91	57.43	27.83	3.75	290.97	0.3	1004.5	2015	7	1	16	WNW	19.32
2015-07-01 17:00:00	14	44	4.07	3.54	0.5	7.73	62.01	26.15	2.69	254.43	0	1004.9	2015	7	1	17	WSW	18.55
2015-07-01 18:00:00	11	44	2.27	3.45	0.44	5.28	60.2	22.92	2.34	270.66	0	1005.4	2015	7	1	18	W	14.96
2015-07-01 19:00:00	8	32	2.66	3.45	0.56	5.35	61.54	23.22	0.96	207.2	0	1006	2015	7	1	19	SSW	15.53
2015-07-01 20:00:00	13	50	2.61	3.41	0.7	5.05	63.7	23.39	1.51	125.22	0	1006.6	2015	7	1	20	SE	16.13
2015-07-01 21:00:00	24	67	2.47	3.31	0.89	4.59	65.86	23.43	1.37	182.2	0	1006.6	2015	7	1	21	S	16.6
2015-07-01 22:00:00	18	57	1.49	3.32	0.43	5.05	66.72	23.4	1.43	194.14	0	1006.7	2015	7	1	22	SSW	16.74
2015-07-01 23:00:00	7	36	1.65	3.34	0.37	4.85	67.15	23.51	1.65	163.56	0	1006.5	2015	7	1	23	SSE	16.94
2015-07-02 00:00:00	12	45	2.31	3.3	0.25	4.99	66.94	23.57	1.73	179.45	0	1005.8	2015	7	2	0	S	16.96
2015-07-02 01:00:00	12	44	2.08	3.22	0.2	4.03	66.36	23.62	1.88	239.84	0	1005.5	2015	7	2	1	WSW	16.89
2015-07-02 02:00:00	16	40	0.74	3.16	0.19	4.86	67.11	23.73	1.73	224.38	0	1005.1	2015	7	2	2	SW	17.15
2015-07-02 03:00:00	10	26	0.68	3.13	0.16	5.11	67.08	23.57	1.28	243.07	0	1004.7	2015	7	2	3	WSW	16.99
2015-07-02 04:00:00	9	29	1.18	3.41	0.18	4.43	68.53	23.57	0.89	313.15	0	1004.3	2015	7	2	4	NW	17.28
2015-07-02 05:00:00	10	51	2.35	3.58	0.2	4.7	69.94	23.49	1.28	267.73	0	1004.7	2015	7	2	5	W	17.48

PM10 Analysis: Distribution of the hourly recorded PM10 concentration in the air in Chandrapur district is represented by the graph below which represent the data that the maximum number of data collected is around 100 or higher value in the data set and it is highly right-skewed as some of the data are very higher. It may be because of some events which lead to producing high pollution content production or due to the error in the sensor which collects the data.



The daily average of PM10 contained in the air with the passing of the year is shown below. It is clearly shown that the increase in the population and demand lead to the production of pollution. We can see from the year 2018 concentration is decreasing by imposing proper regulation and rules which are followed by the companies. Together effort has lead in declining in PM10 concentration**.**

Splitting the dataset:

In this process, we split the dependent variable and independent variable so which is the respective feature vector and which helps in the process of machine learning. We split the data into predictor variables and an Outcome variable.

The data set which is to be split depends on the basis is plot by the sns package of available and it helps in the visualization of the data set with each other and how all variable effect on the other variable.

The below table is the correlation data table which describe same as the above figure givesthe detail.

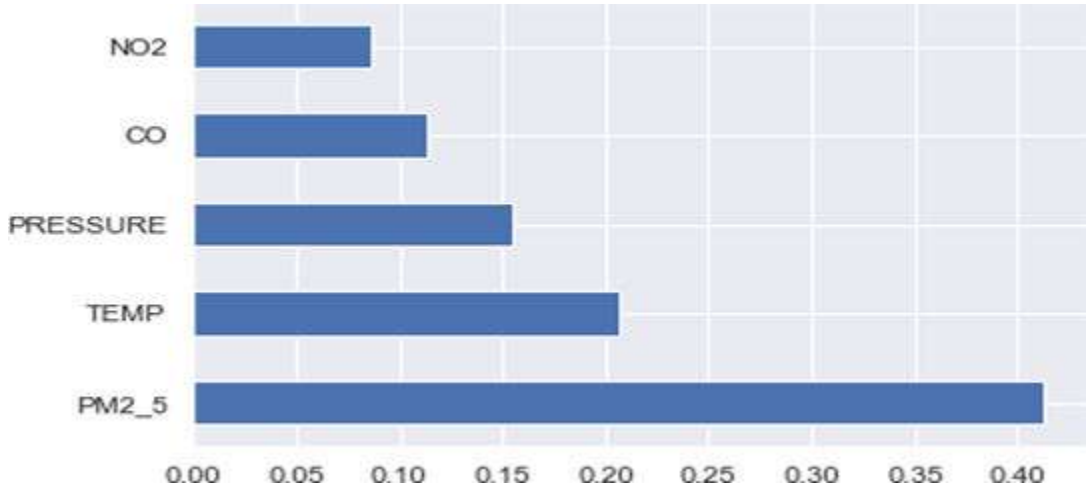
Sr.No	PM2.5	PM10	NO2	CO	TEMP	RAIN	PRESSURE
PM2.5	1.00	0.5912	0.106	0.3867	-0.1344	-0.0858	0.3767
PM10	0.5912	1.00	0.1244	0.4037	-0.12267	-0.1480	0.3582
NO2	0.1061	0.1244	1	0.1244	0.3037	0.0029	0.0789
CO	0.386	0.4030	0.12148	1.00	-0.1672	-0.0613	0.2118
TEMP	-0.1344	-0.2267	0.30377	-0.167	1.000	0.0540	-0.0364
RAION	-0.085	0.14809	0.0029	-0.061	0.0540	1.000	-0.1962
PRESSURE	0.3767	0.3582	0.0789	0.211	-0.0363	-0.1962	1.000

Feature Selection:

After processing the given code we can get the feature impotence value which explains theimportance.

[0.41243796 0.08635037 0.11407872 0.20712723 0.02498067 0.15502504]

We can plot graph of feature importances for better visualization which gives more detail.



Feature Importance Image

This function provides access to several approaches for visualizing the univariate of Applying the different machine learning model to the above final data

Linear Regression: Code for processing linear regression model with their output. Here we have imported the linear regression model from sklearn package. We haveuse the fit command which is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') andone or more independent variables (often called 'predictors', 'covariates', or 'features').

We can see that R^2 Value obtained on the train and testing data set are 0.4345 and 0.4223 respectively. After doing that we can see the model evaluation part which tells us:-

1. Holding all other features fixed, a 1 unit increase in PM2_5 is associated with an increase of 0.008 in AQI PM10.
2. Holding all other features fixed, a 1 unit increase in NO2 is associated withan increase of 0.009in AQI PM 10.
3. Holding all other features fixed, a 1 unit increase in CO is associated withan increase of 0.33 inAQI PM 10.
4. Holding all other features fixed, a 1 unit increase in TEMP is associated with a decrease of 0.005 in AQI PM 10.
5. Holding all other features fixed, a 1 unit increase in RAIN is associatedwith a decrease of0.07544 in AQI PM 10.
6. Holding all other features fixed, a 1 unit increase in PRESURE isassociated with an increase of0.0166 in AQI PM 10.

RESULT AND CONCLUSION

So now we know which model perform better in this type of data. On the basis of past data the air quality in Chandrapur worsened between 2015 and 2021 and PM10 concentrations increased. We can predict the Air Pollution Content present in the environment like PM10 and it will be helping in the monitoring of Air Quality Index. It will be very helpful in maintaining the health of human beings and leaving organism and historical monuments which get damage by the content present in the e environment. The Air pollution modeling using machine learning isproving to be a powerful technique in determining and investigating changes in emissions that lead to improved, or deteriorated, air quality. This study used linear regression, Lasso regression, Decision Tree regression, Random forest regression, K-Nearest neighbor, XGboost regression and ANN models to investigate the changes in PM10 in Chandrapur. By accounting for meteorological and environmental effects, the results show

that there has been an increase per year due to changes in local emissions, with the remainder mostly due to a decrease in soil water content in the region which can facilitate dust emissions.

The reported PM10 emissions from each individual mine site surrounding Chandrapur as well as the total emissions from all sites and the total for different directions. This study extended on the meteorological dataset applied in a proxyfor fire emissions, which have significant impacts on air quality.

REFERENCE

1. Marc Daniel Mallet, Atmospheric Pollution Research 12 (2021) 23– 35, "Meteorological normalization of PM10 using machine learning reveals distinct increases of nearby source emissions in the Australian mining town of Moranbah".
2. Chaulya, S.K.,2005,— "Air Quality status of an Open Pit Mining Area in India", Environment Monitoring and Assessment, Vol. 105, pp. 369 – 389
3. Chinthala Sumanth and Khare M. 2011 , "Particle Dispersion within a Deep Open Cast Mine, Air Quality Models and Applications", Prof. Dragana Popovic (Ed.) ISBN 978-953-307-307-1, 280-289
4. S. Jeya, Dr. L. Sankari, "Air Pollution Prediction by Deep Learning Model", Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
5. A.Suleiman, M.R. Tight, A.D. Quinn, "Applying machine learning methods in managing urban concentrations of traffic-related particulatematter (PM10 and PM2.5)", Atmospheric Pollution Research
6. Trivedi R., Chakraborty M.K., Tewary B.K.,2008,— A Study Dust Dispersion of Airborne Dust Generated due to Mining Activities in